Regresión Lineal Múltiple

Bibliografía:

- Chaterjee, S.; Hadi, A.; Price, B. "Regression Analysis by Example". Wiley
- Montgomery, D.; Peck, E.; Vining, G. "Introducción al Análisis de Regresión Lineal". 3a ed.
- Draper, N.R.; Smith H. (1981) "Applied Regression Analysis". 2nd ed. Wiley N.Y.

Guía de ruta:

- Modelo RLM
- Estimación de mínimos cuadrados
- Supuestos del modelo RLM
- Inferencia en el modelo RLM
- Verificando los supuestos
- Diagnósticos de influencia
- Multicolinealidad
- RLM con variables dummies
- Métodos de selección de variables
- Transformaciones

Variables dummies

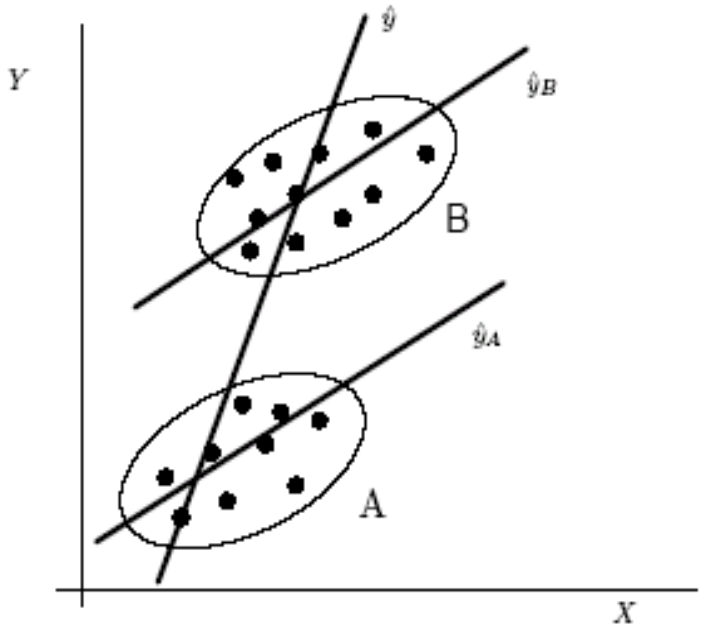
En un muestra pueden aparecer grupos de observaciones. El modelo de regresión lineal puede no ser adecuado si no contempla la existencia de estos grupos.

Ejemplo:

Y=peso

X= altura

Tenemos datos de distinto sexo.



Variables dummies

Para considerar la posibilidad de un ajuste diferente en cada grupo podemos introducir variables *ficticias*, *dicotómicas* o *dummies* del siguiente modo:

$$z_i = \begin{cases} 0 \text{ si la observación } i \text{ pertenece al grupo A} \\ 1 \text{ si la observación } i \text{ pertenece al grupo B} \end{cases}$$

Regresión con una sola variable dummy

Consideremos un modelo de regresión con una sola variable dummy Z y una variable cuantitativa X. Es decir,

$$Y=\beta_0+\beta_1 X+\delta Z + \varepsilon$$

Entonces el modelo considerado en cada grupo es:

Si $Z=0$, $Y= \beta_0 + \beta_1 X + \varepsilon$

Si $Z=1$, $Y= (\beta_0 + \delta) + \beta_1 X + \varepsilon$

Es decir que el modelo **considera que las pendientes de ambas lineas son iguales.**

El valor estimado de δ representa el cambio promedio en la variable de respuesta al cambiar el valor de la variable "dummy".

Ejemplo: tipo de herramienta ¿Cómo son los datos?

18,73	610,00	A	0
14,52	950,00	A	0
17,43	720,00	A	0
14,54	840,00	A	0
13,44	980,00	A	0
24,39	530,00	A	0
13,34	580,00	A	0
22,71	540,00	A	0
12,68	890,00	A	0
19,32	730,00	A	0
30,16	670,00	B	1
27,09	770,00	B	1
25,40	880,00	B	1
26,05	1000,00	B	1
33,49	760,00	B	1
35,62	590,00	B	1
26,07	910,00	B	1
36,78	650,00	B	1
34,95	810,00	B	1
43,67	500,00	B	1

Para nuestro ejemplo:

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	35,21	3,74	27,32	43,10	9,42	<0,0001
xi1 (rpm)	0,02	4,9E-03	-0,03	-0,01	-5,05	0,0001
Tool Type_B	15,24	1,50	12,07	18,40	10,15	<0,0001

nuestro modelo es

$$\hat{y} = 35.21 + 0.02x + 15.24z$$

el coeficiente de z es significativamente diferente de cero, entonces….¿cómo es el modelo que ajustamos a cada grupo?

Si hay más de 2 grupos, ¿cuántas dummies?

Necesitamos (k-1) variables dummies si tenemos k categorías. Por ejemplo con 3 categorías:

$$Y=\beta_0+\beta_1 X+\delta_1 Z_1 +\delta_2 Z_2 + \varepsilon$$

entonces:

Si $Z_1=0, Z_2=0$: $Y= \beta_0 + \beta_1 X + \varepsilon$

Si $Z_1=0, Z_2=1$: $Y=\beta_0+\beta_1 X+ \delta_2 + \varepsilon$

Si $Z_1=1, Z_2=0$: $Y=\beta_0+\beta_1 X+ \delta_1 + \varepsilon$

Es decir que el modelo **considera también que las pendientes de las lineas son iguales.**

Ejemplo: con 3 grupos y 1 predictor continuo, el modelo anterior permite evaluar si

a) Las lineas son iguales: o sea no hay diferencias entre las medias de los grupos ($\delta_1 = \delta_2 = 0$)

b) La línea de un grupo no difiere de la del grupo base, o sea la media del grupo "i" no difiere de la media del grupo base ($\delta i = 0$)

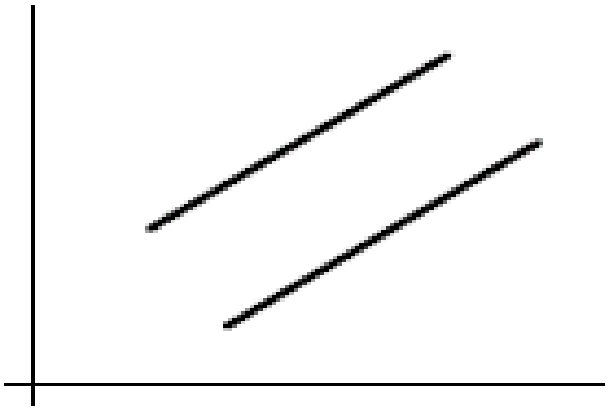
c) etc..

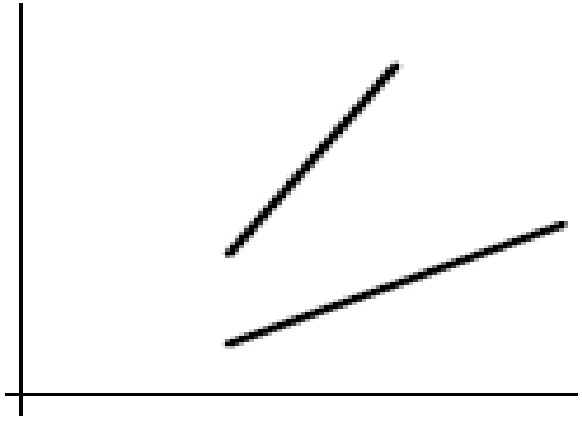
Pero este modelo no toma en cuenta la posibilidad de pendientes distintas!

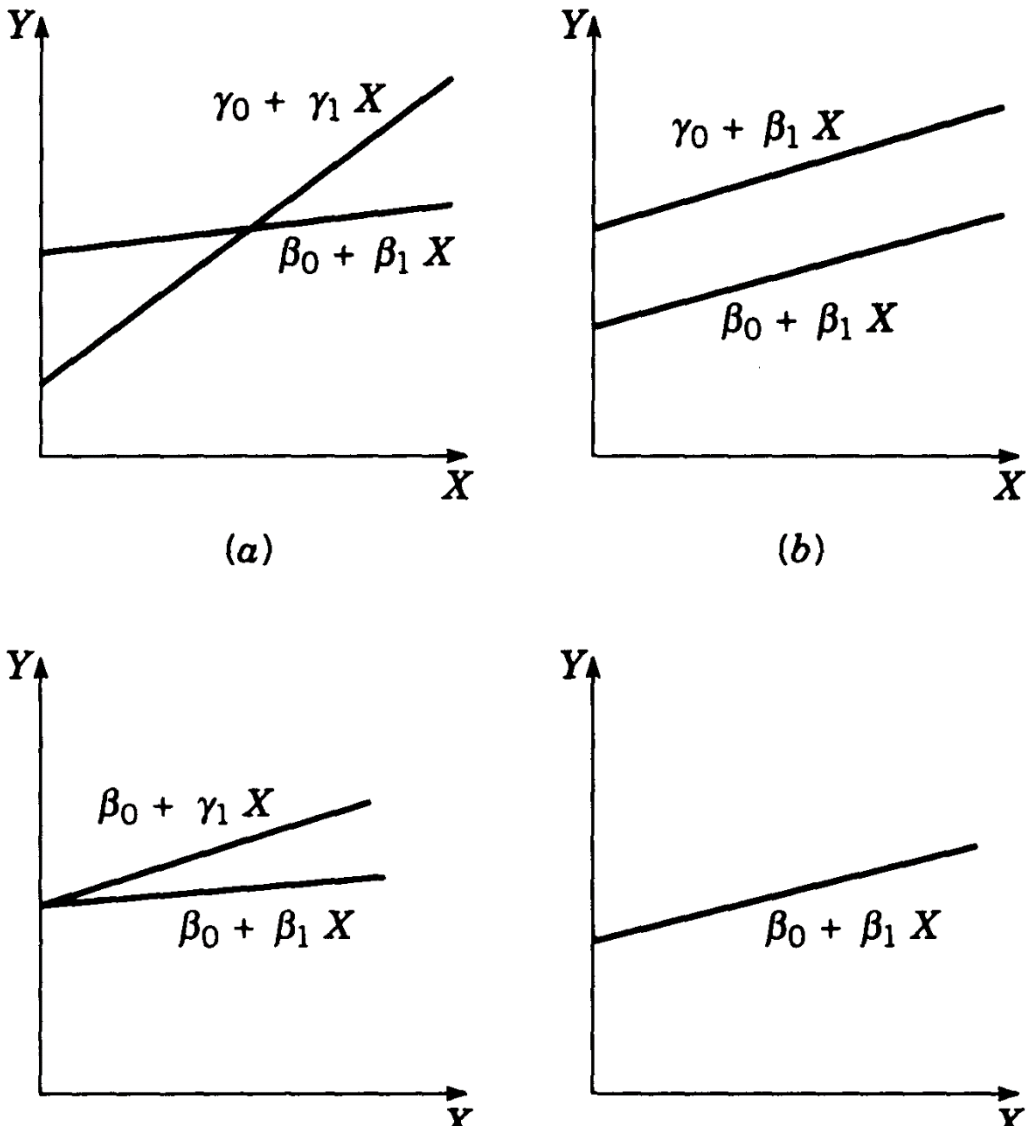
Interacción

Decimos que hay **interacción** entre dos variables cuando la variable producto de ambas tiene efecto significativo.
Esta variable ficticia permite modelar el efecto de una de las variables sobre la respuesta, dependiendo de la otra variable.
Esto permite distintas pendientes por categoría en la regresión.

$$y = \beta_0 + \beta_1 x + \delta z + \varepsilon \qquad y = \beta_0 + \beta_1 x + \delta z + \gamma xz + \varepsilon$$






Y
$\gamma_0 + \gamma_1 X$
$\beta_0 + \beta_1 X$
X
(a)
Y
$\gamma_0 + \beta_1 X$
$\beta_0 + \beta_1 X$
X
(b)
Y
$\beta_0 + \gamma_1 X$
$\beta_0 + \beta_1 X$
Y
Y
$\beta_0 + \beta_1 X$
Y

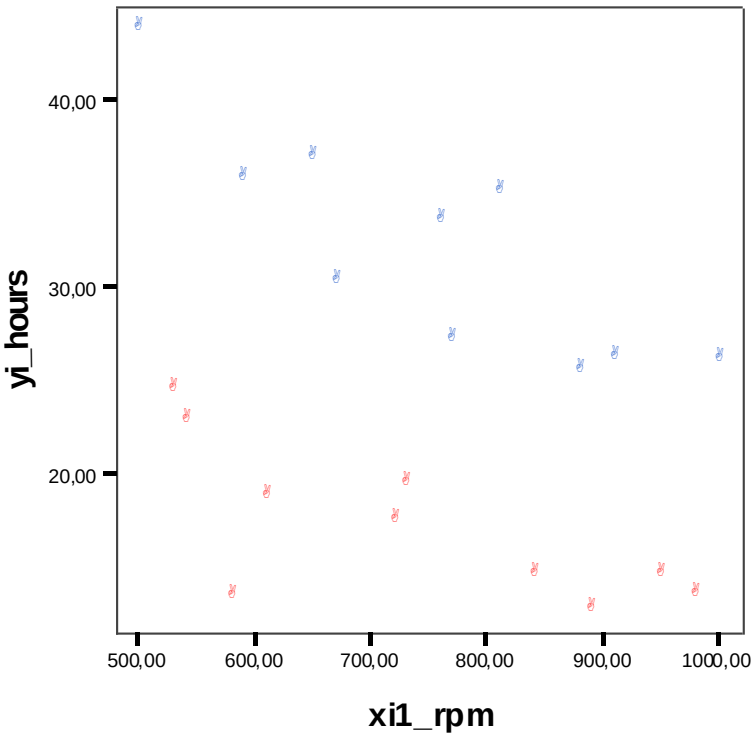
(de Draper)

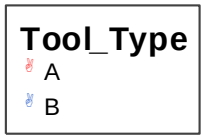
Test de hipótesis de interés en este modelo

- Las pendientes son iguales en ambos grupos (o sea no hay interacción), corresponde a $H_0 : \gamma = 0$
- Las líneas son iguales, o sea no hay diferencias entre las medias de los grupos ($\delta = \gamma = 0$)

Se usan pruebas F de modelo completo versus modelo reducido como planteamos antes en rlm.

Ejemplo: tiempo de duración de herramientas de tipos A y B (ej 8.2 Montgomery)





Los puntos/líneas muestran Medias

MOD 1

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,937[a]	,879	,864	3,35173

a. Variables predictoras: (Constante), tipo, xi1_rpm

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1384,108	2	692,054	61,603	,000[a]
	Residual	190,980	17	11,234		
	Total	1575,088	19			

a. Variables predictoras: (Constante), tipo, xi1_rpm

b. Variable dependiente: yi_hours

Tipo=0 si A

Tipo=1 si B

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	35,209	3,739		9,417	,000
	xi1_rpm	-,025	,005	-,427	-5,048	,000
	tipo	15,235	1,501	,858	10,149	,000

a. Variable dependiente: yi_hours

Hay interacción???

MOD 2

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	56,745	5,321		10,665	,000
	xi1_rpm	-,018	,006	-,308	-2,831	,012
	tipoH	-26,569	7,116	-1,497	-3,734	,002
	Xtipo	-,015	,009	-,669	-1,626	,123

a. Variable dependiente: yi_hours

Cuál es el modelo ajustado para cada tipo de herramienta??

Datos: (Chaterjee- pag 145)

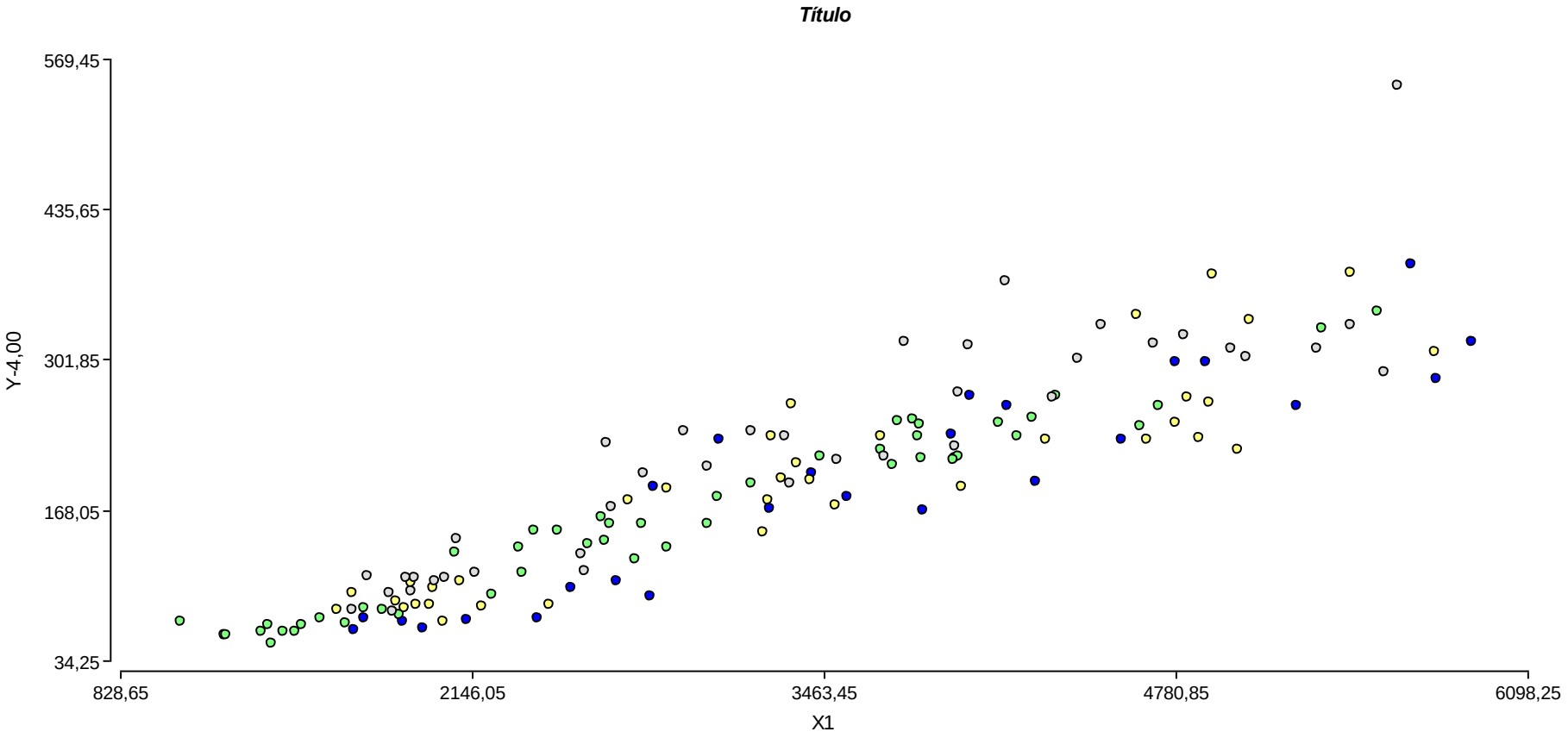
Y : 'gastos' per capita en educación en 1960

X1: ingresos per capita

X2: proporción de la población por debajo de 18 años (por mil)

X3: proporción de la población que reside en áreas urbanas (por mil)

R: región (4 posibles)


Título
569,45
435,65
301,85
168,05
34,25
Y-4,00
828,65
2146,05
3463,45
4780,85
6098,25
X1

State	Y	X1	X2	X3	Region	R1	R3	R4
ME	61	1704	388	399	1	1	0	0
NH	68	1885	372	598	1	1	0	0
VT	72	1745	397	370	1	1	0	0
….								
OH	82	2184	387	674	2	0	0	0
IN	84	1990	392	568	2	0	0	0
IL	84	2435	366	759	2	0	0	0
…								
DE	124	2760	388	326	3	0	1	0
MD	92	2221	393	562	3	0	1	0
….								
MT	95	1920	412	463	4	0	0	1
ID	79	1701	418	414	4	0	0	1
WY	142	2088	415	568	4	0	0	1
CO	108	2047	399	621	4	0	0	1

Modelo 1: con sólo X1 (región 2 como base) :

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	-44,44	9,79	-63,79	-25,08	-4,54	<0,0001
X1	0,07	2,3E-03	0,06	0,07	29,68	<0,0001
Region_1,00	-14,33	8,81	-31,74	3,09	-1,63	0,1062
Region_3,00	6,14	7,73	-9,13	21,42	0,80	0,4277
Region_4,00	33,07	7,99	17,28	48,86	4,14	0,0001

Modelo con interacciones

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor
const	-32,94	16,80	-66,14	0,27	-1,96	0,0518
X1	0,07	4,6E-03	0,06	0,07	14,17	<0,0001
Region_1,00	-23,51	26,32	-75,54	28,51	-0,89	0,3731
Region_3,00	2,78	21,37	-39,46	45,01	0,13	0,8967
Region_4,00	-1,65	23,37	-47,85	44,56	-0,07	0,9439
Region_1,00 X1	2,7E-03	0,01	-0,01	0,02	0,39	0,6998
Region_3,00 X1	5,3E-04	0,01	-0,01	0,01	0,09	0,9324
Region_4,00 X1	0,01	0,01	-2,5E-03	0,02	1,58	0,1157

Como se muestra en rojo, no son significativas las interacciones por lo que se descartan y se prefiere el modelo sin estas (Mod 1). Esto es, las rectas para cada región no tienen pendientes significativamente diferentes.

Técnicas automáticas para selección de variables

Una aplicación de los test de hipótesis para los coeficientes lo dan los métodos "automáticos"de selección de variables.

Estos tratan de elegir un modelo que explique el comportamiento de la variable respuesta lo mejor posible, haciendo uso del menor número de variables predictoras posibles, esta propiedad es llamada *"parsimonia"*.

Hay situaciones en que consideraciones teóricas determinan la elección de variables a incluir. Pero también existen métodos automáticos para lograr este objetivo: los métodos *backward*, *forward y stepwise*.

Metodos paso a paso

La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo (o excluyendo) una variable predictora en cada paso de acuerdo a ciertos criterios.

El proceso secuencial termina cuando una regla de parada se satisface.

Tres algoritmos muy usados para seleccionar variables son:

- Backward Elimination
- Forward Selection
- Stepwise Selection

Backward Elimination

- Se comienza con el modelo completo y en cada paso se va eliminando una variable.

- Si resultara que todas las variables predictoras son importantes, es decir, tienen “p-value” pequeños para la prueba t, entonces no se hace nada y se concluye que el mejor modelo es el que tiene todas las variables predictoras disponibles.

Backward Elimination

En cada paso la variable que se elimina del modelo es aquella que satisface cualquiera de estos requisitos equivalentes entre sí:

- Aquella variable que tiene el estadístico de *t (*en valor absoluto) más pequeño entre las variables incluidas aún en el modelo (o sea mayor *p-valor*)
- Aquella variable que produce la menor disminución en el R^2 al ser eliminada del modelo.

Se para cuando el mínimo de los valores *t* es $>(n-p)*t_{0.05}$

o mayor que 2. (similar a un F_{out})

Ejemplo: datos de cemento de Hald

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2667,899	4	666,975	111,479	,000[a]
	Residual	47,864	8	5,983		
	Total	2715,763	12			

a. Variables predictoras: (Constante), x4, x3, x1, x2

.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%		Estadísticos de colinealidad	
		B	Error típ.	Beta			Límite inferior	Límite superior	Tolerancia	FIV
1	(Constante)	62,405	70,071		,891	,399	-99,179	223,989		
	x1	1,551	,745	,607	2,083	,071	-,166	3,269	,026	38,496
	x2	,510	,724	,528	,705	,501	-1,159	2,179	,004	254,4
	x3	,102	,755	,043	,135	,896	-1,638	1,842	,021	46,868
	x4	-,144	,709	-,160	-,203	,844	-1,779	1,491	,004	282,5

a. Variable dependiente: y

En el modelo completo, todos las regresoras, se ven signos multicolinealidad

Seleccionando con backward

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%		Estadísticos de colinealidad	
		B	Error típ.	Beta			Límite inferior	Límite superior	Tolerancia	FIV
1	(Constante)	62,405	70,071		,891	,399	-99,179	223,989		
	x1	1,551	,745	,607	2,083	,071	-,166	3,269	,026	38,496
	x2	,510	,724	,528	,705	,501	-1,159	2,179	,004	254,423
	x3	,102	,755	,043	,135	,896	-1,638	1,842	,021	46,868
	x4	-,144	,709	-,160	-,203	,844	-1,779	1,491	,004	282,513
2	(Constante)	71,648	14,142		5,066	,001	39,656	103,641		
	x1	1,452	,117	,568	12,410	,000	1,187	1,717	,938	1,066
	x2	,416	,186	,430	2,242	,052	-,004	,836	,053	18,780
	x4	-,237	,173	-,263	-1,365	,205	-,629	,155	,053	18,940
3	(Constante)	52,577	2,286		22,998	,000	47,483	57,671		
	x1	1,468	,121	,574	12,105	,000	1,198	1,739	,948	1,055
	x2	,662	,046	,685	14,442	,000	,560	,764	,948	1,055

a. Variable dependiente: y

Forward Selection

- Se empieza con la regresión lineal simple que considera como variable predictora a aquella que está más altamente correlacionada con la variable de respuesta.

- Si esta primera variable no es significativa entonces se reconsidera este modelo y se para el proceso.

Forward Selection

Si hay variables que son significativas se añade al modelo la variable que reune cualquiera de estos requisitos equivalentes entre sí:

- Aquella variable que tiene el estadístico de *t* (en valor absoluto) más grande entre las variables no incluidas aún en el modelo. Es decir, la variable con el *F-parcial* más grande.

- Aquella variable que produce el mayor incremento en el R^2 al ser añadida al modelo. Es decir, aquella variable que produce la mayor reducción en la suma de cuadrados del error.

Criterios de parada para el metodo forward

- El valor de la prueba de *F parcial* para cada una de las variables no incluidas aún en el modelo es menor que un número prefijado *F-in* (por lo general este valor es 4).

- Cuando el valor absoluto del estadistico de t es menor que la raíz cuadrada de *F-in* (por lo general, $|t|<2$).

Seleccionando con forward

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%		Estadísticos de colinealidad	
		B	Error típ.	Beta			Límite inferior	Límite superior	Tolerancia	FIV
1	(Constante)	117,57	5,262		22,3	,000	105,986	129,150		
	x4	-,738	,155	-,821	-4,77	,001	-1,078	-,398	1,000	1,000
2	(Constante)	103,10	2,124		48,5	,000	98,365	107,830		
	x4	-,614	,049	-,683	-12,6	,000	-,722	-,506	,940	1,064
	x1	1,440	,138	,563	10,4	,000	1,132	1,748	,940	1,064

a. Variable dependiente: y

Stepwise Selection

- Se puede considerar como una modificación del método "Forward". Es decir, se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada. Aqui se usan *F-out* y *F-in* con $F\text{-}in \leq F\text{-}out$.

- El proceso termina cuando ninguna de las variables, que no han entrado aún, tienen importancia suficiente como para entrar al modelo.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	57,424	8,491		6,763	,000
	x2	,789	,168	,816	4,686	,001

a. Variable dependiente: y

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	94,160	56,627		1,663	,127
	x2	,311	,749	,322	,415	,687
	x4	-,457	,696	-,508	-,657	,526

a. Variable dependiente: y

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Estadísticos de colinealidad	
		B	Error típ.	Beta			Tolerancia	FIV
1	(Constante)	203,642	20,648		9,863	,000		
	x2	-,923	,262	-,955	-3,525	,006	,041	24,309
	x4	-1,557	,241	-1,732	-6,454	,000	,042	23,859
	x3	-1,448	,147	-,616	-9,846	,000	,770	1,298

a. Variable dependiente: y

Multicolinealidad

Correlaciones

		y	x1	x2	x3	x4
y	Correlación de Pearson	1	,731**	,816**	-,535	-,821**
	Sig. (bilateral)		,005	,001	,060	,001
	N	13	13	13	13	13
x1	Correlación de Pearson	,731**	1	,229	-,824**	-,245
	Sig. (bilateral)	,005		,453	,001	,419
	N	13	13	13	13	13
x2	Correlación de Pearson	,816**	,229	1	-,139	-,973**
	Sig. (bilateral)	,001	,453		,650	,000
	N	13	13	13	13	13
x3	Correlación de Pearson	-,535	-,824**	-,139	1	,030
	Sig. (bilateral)	,060	,001	,650		,924
	N	13	13	13	13	13
x4	Correlación de Pearson	-,821**	-,245	-,973**	,030	1
	Sig. (bilateral)	,001	,419	,000	,924	
	N	13	13	13	13	13

**. La correlación es significativa al nivel 0,01 (bilateral).

Transformaciones en RLM

Pueden ser necesarias por diversas razones:

- La relación no es lineal entre X e Y, desde consideraciones teóricas u observación de los datos.
- La varianza de Y no es homogénea, depende de la media (o sea Var(y) cambia con X)
- Al examinar residuales hay indicios de heterogeneidad o falta de normalidad.

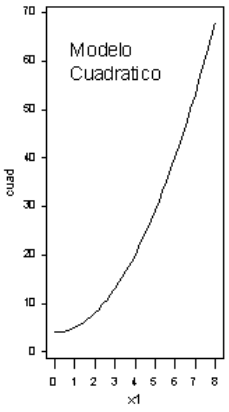
Transformaciones para linealizar modelos

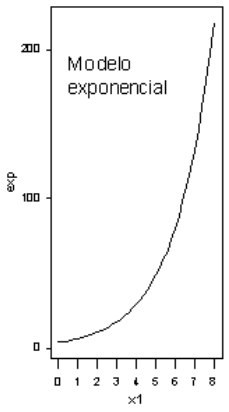
Consideremos por ahora solo modelos con una sola variable predictora.

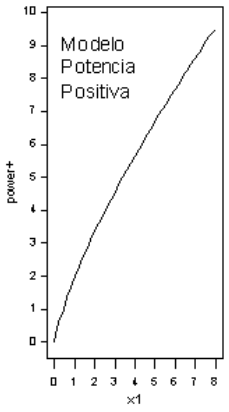
El objetivo es tratar de transformar las variables para mejorar el ajuste del modelo, sin incluir variables predictoras adicionales.

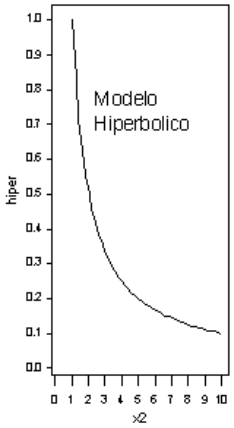
Transformaciones de la variable predictora y/o respuesta para linealizar varios modelos.

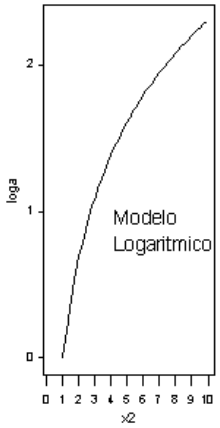
Nombre del modelo	Ecuación del Modelo	Transformación	Modelo Linealizado
Exponencial	$Y=\alpha e^{\beta X}$	$Z=\text{Log}Y \quad X=X$	$Z=\text{Log}\,\alpha+\beta X$
Logaritmico	$Y=\alpha+\beta\text{Log}\,X$	$Y=Y \quad W=\text{Log}\,X$	$Y=\alpha+\beta W$
Doblemente Logarítmico o Potencia	$Y=\alpha X^{\beta}$	$Z=\text{Log}\,Y \quad W=\text{Log}\,X$	$Z=\text{Log}\,\alpha+\beta W$
Hiperbólico	$Y=\alpha+\beta/X$	$Y=Y \quad W=1/X$	$Y=\alpha+\beta W$
Doblemente Inverso	$Y=1/(\alpha+\beta X)$	$Z=1/Y \quad X=X$	$Z=\alpha+\beta X$

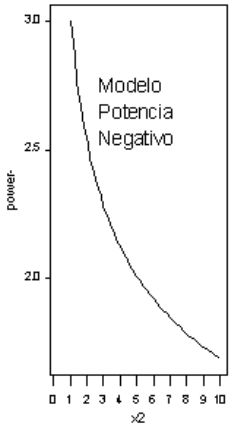












Algunas transformaciones en la variable respuesta para estabilizar la varianza.

Transformación	**Situación**
$\sqrt{y}$	$Var(e_i) \propto E(y_i)$
$\sqrt{y} + \sqrt{y+1}$	$Var(e_i) \propto E(y_i)$
$Log(Y)$	$Var(e_i) \propto (E(y_i))^2$
$Log(y+1)$	$Var(e_i) \propto (E(y_i))^2$
$1/y$	$Var(e_i) \propto (E(y_i))^4$

Ejemplo:

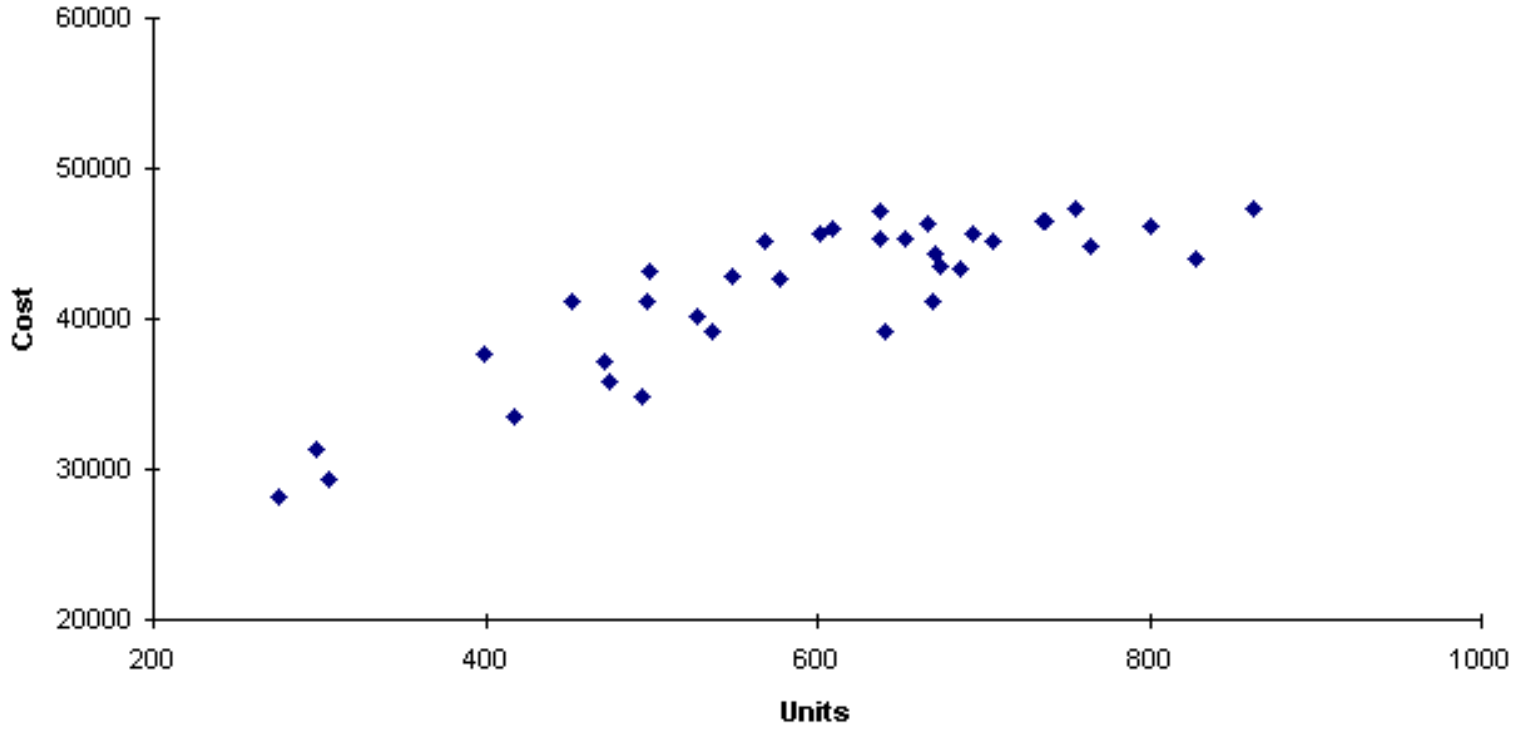
- Una empresa de energía produce diferentes cantidades por mes, dependiendo de la demanda .
- Datos: unidades producidas por (Units) y costo total (Cost) en un período de 36 meses.
- Cómo podemos usar una regresión para analizar la relación entre Cost y Units?

Algunos datos

cost versus production level		
Cost	Units	
45623	601	
46507	738	
43343	686	
46495	736	
47317	756	
41172	498	
43974	828	
44290	671	
29297	305	
47244	637	
46295	667	
45218	705	
45357	637	

Scatter plot

Antes que nada: gráfico de los datos



A simple vista:

Se puede ver una relación creciente..¿lineal?

Más parece una cuadrática, pero veamos primero un ajuste lineal.

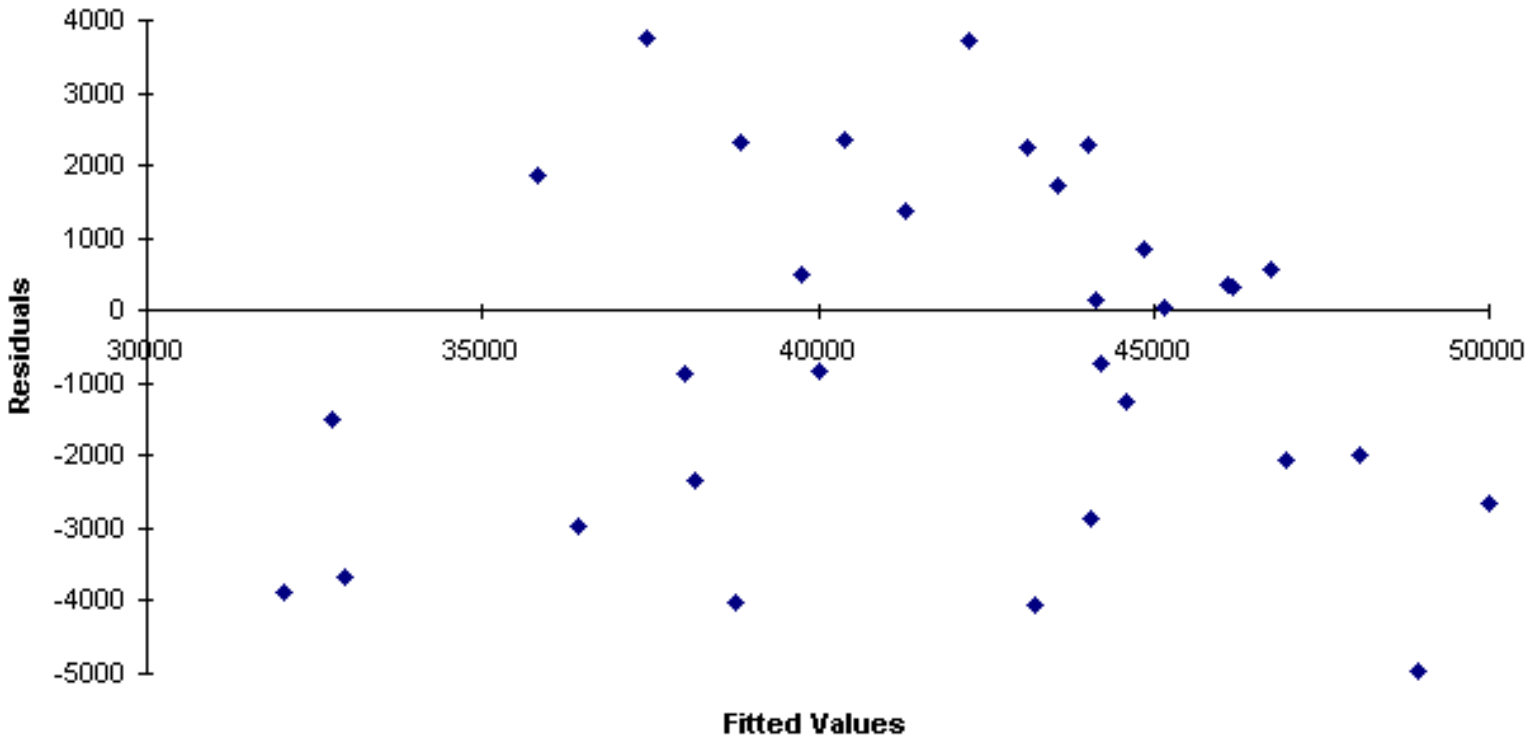
Solución ajustada

- La ecuación ajustada es

  Cost^ = 23,651 + 30.53 Units

- R^2 =73.6%
- Graficamos residuales para ver supuestos:

Residuales vs predichos



Qué se puede observar?

- Puede verse un patrón no lineal
- Sugiere una parábola, entonces creamos la variable Units^2

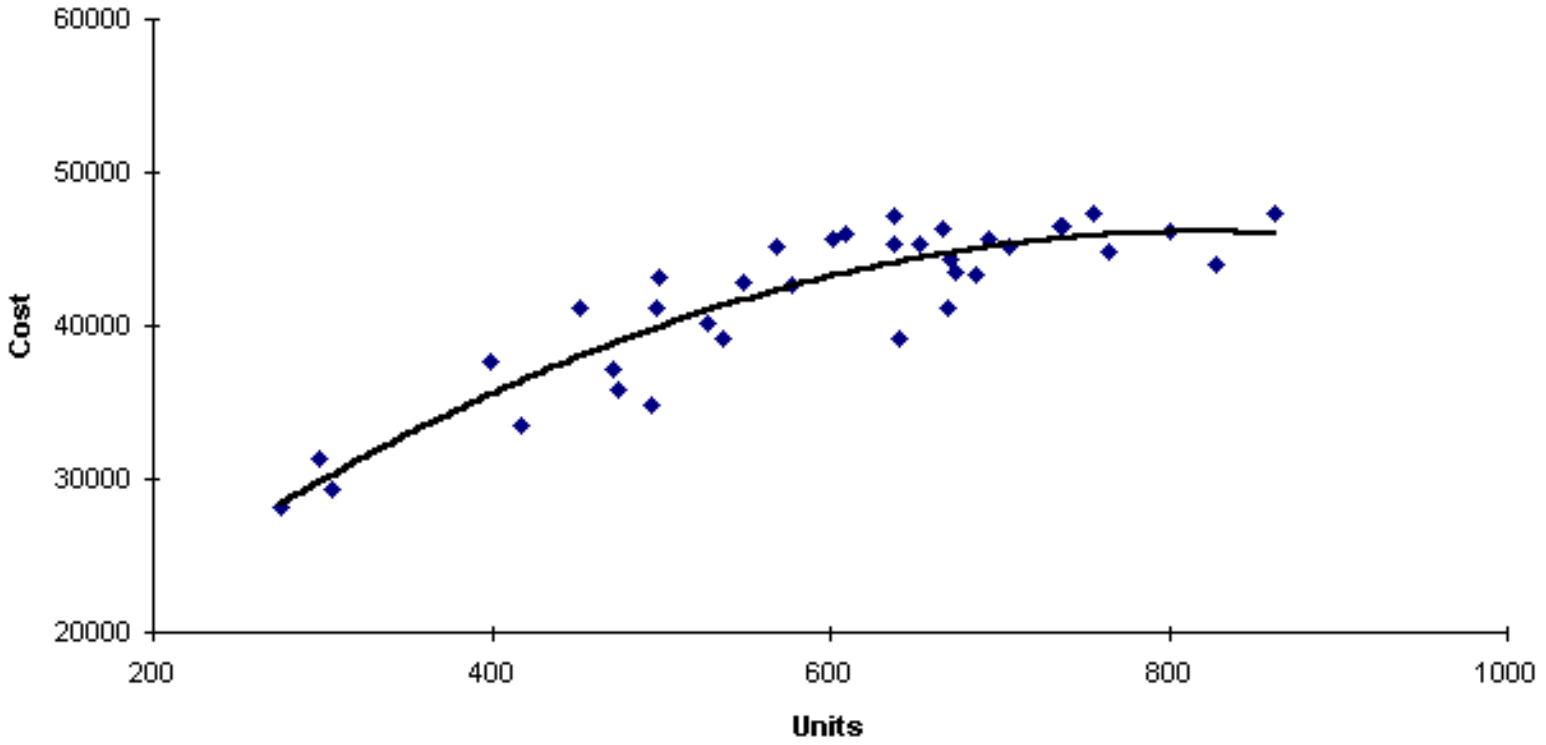
Ajuste de regresión

La ecuación resultante

Cost^ = 5793 +98.3Units - 0.0600Units^2

R^2 = 82.2%

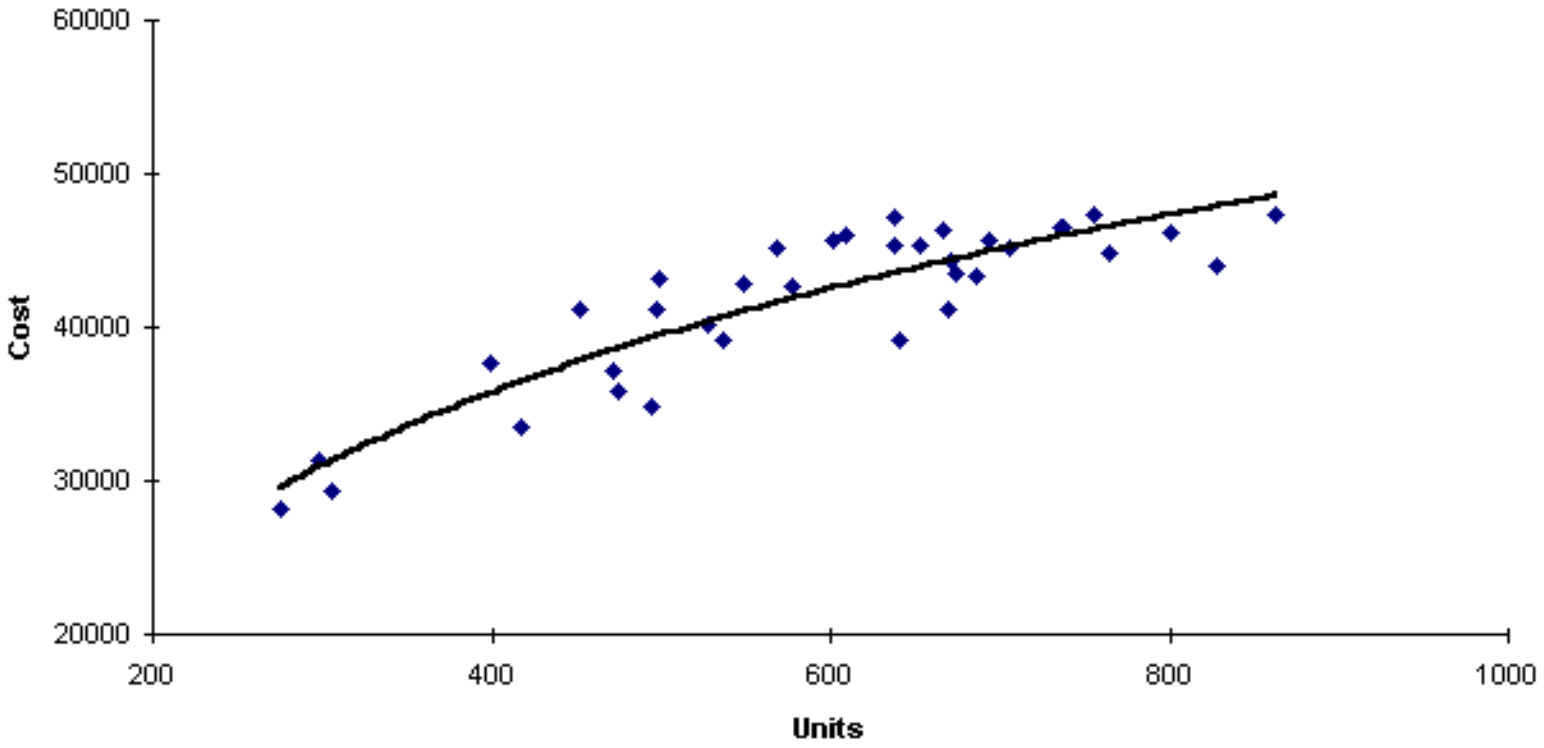
Dibujo la curva ajustada sobre los datos



- El problema en este tipo de modelos es la interpretación de los coeficientes, especialmente del término cuadrático.
- Otra posibilidad: ajuste logarítimico.

  Creamos la variable LogUnits.

Ajuste Logarítmico



Modelo ajustado

Cost^ = -63,993 + 16,654LogUnits

R^2 = 79.8%

No son mejores que el ajuste cuadrático, pero la ventaja del modelo logarítmico es que es más fácil de interpretar: un aumento del 10% en Units provoca que el costo aumente 16.654*log(1.1)=1.5873

Selección de variables en RLM

- Hasta ahora:

supusimos que todas las variables regresoras incluídas eran relevantes >> evaluamos si la forma del modelo era correcta, supuestos se cumplían, no había puntos influyentes, no colinealidad, etc.

- Sin embargo, en la realidad la situación es:

se tienen variables *candidatas* >> debemos elegir qué subconjunto de ellas dan un modelo adecuado.

Cómo? Con métodos para seleccionar modelos.

- Luego, chequeamos supuestos, influyentes, etc., lo cual podría llevar nuevamente a reelegir el conjunto de variables a incluír.

Consecuencias de la mala especificación del modelo (cap. 9 Montgomery)

- Los estimadores de los coeficientes pueden dar sesgados.

$$Y = X_p \beta_p + X_r \beta_r + \varepsilon^* \quad \text{Modelo completo}$$

$$Y = X_p \beta_p + \varepsilon \quad \text{Modelo reducido}$$

$$\hat{\beta}^* = \left(X' X \right)^{-1} X' Y \quad \text{Estimacion en el Modelo completo}$$

$$\hat{\beta}_p = \left(X_p' X_p \right)^{-1} X_p' Y \quad \text{Estimacion en el Modelo reducido}$$

$$E\left[\hat{\beta}_p \right] = \beta_p + \underbrace{\left(X_p' X_p \right)^{-1} X_p' X_r}_{A} \beta_r$$ por lo que en general estimar

con el modelo chico da sesgado

Consecuencias de la mala especificación del modelo (cap. 9 Montgomery)

- Las varianzas de los estimadores de coef. en el modelo chico no dan mayores que en el grande.

$$Var\left(\hat{\beta}_p^{\,*}\right) = \sigma^2 \left[\left(X'X\right)^{-1}\right]_{(\textit{parte correspondiente})}$$ en el Modelo completo

$$Var\left(\hat{\beta}_p\right) = \sigma^2 \left(X_p' X_p\right)^{-1}$$ en el Modelo reducido

se prueba que

$$Var\left(\hat{\beta}_p^{\,*}\right) - Var\left(\hat{\beta}_p\right) \geq 0$$

Consecuencias de la mala especificación del modelo (cap. 9 Montgomery)

- Las predicciones en el modelo chico dan sesgadas pero con menor error cuadrático medio, lo que da mejor precisión.

$\hat{y}^* = x'_p \hat{\beta}^*_p + x'_r \hat{\beta}^*_r$ prediccion con modelo completo

$\hat{y} = x'_p \hat{\beta}_p$ prediccion con modelo reducido

$\mathrm{E}\left[\hat{y}^*\right] = x'\beta \quad \wedge \quad \mathrm{V}\left[\hat{y}^*\right] = \sigma^2 x' \left(X' X\right)^{-1} x$

$\mathrm{E}\left[\hat{y}\right] = x'_p \beta_p + x'_p A \beta_r$ sesgada!

$\mathrm{V}\left[\hat{y}\right] = \sigma^2 x'_p \left(X'_p X_p\right)^{-1} x_p$

Consecuencias de la mala especificación del modelo (cap. 9 Montgomery)

Lo anterior muestra que hay que comparar ambas predicciones con error cuadrático medio y se tiene que:

ECM de predicción en el modelo completo $\geq$ ECM de predicción en el modelo reducido

Esto motiva la elección de modelos 'más chicos'. Se introduce sesgo pero si el efecto es moderado, se gana en precisión.

Criterios para evaluar modelos

- El coeficiente de determinación R^2
- El R^2 ajustado
- La varianza estimada del error (MSE).
- PRESS (Suma de cuadrados de Predicción)
- C_p de Mallows.

El coeficiente de Determinación R^2

- Se elige el modelo que tenga un R^2 bastante alto con el menor número de variables predictoras posibles.
- Se elige un modelo con k variables si al incluir una variable adicional el R^2 no se incrementa sustancialmente

Algunos problemas de este criterio

- Efecto de datos anormales.
- Agregando variables siempre se aumenta R^2 entonces..¿cuántas agrego?

El R^2 ajustado

Para subsanar la tendencia del R^2 se ha definido un ***R^2-ajustado***

El modelo que se busca es aquel que tiene un ***R^2-ajustado*** alto con pocas variables.

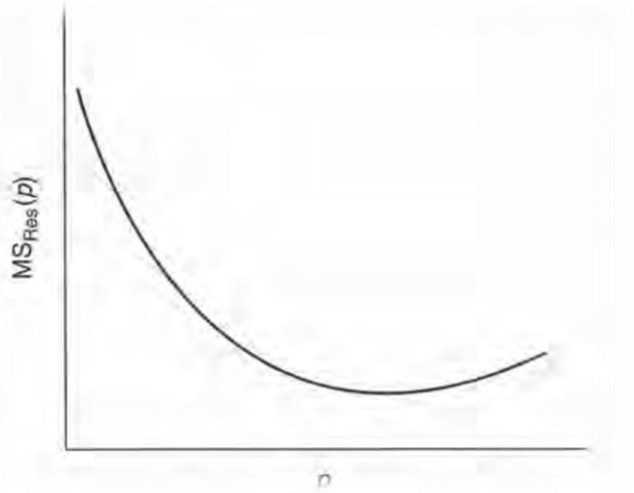
Problemas:

El R^2 ajustado podría disminuir al incluirse una variable adicional en el modelo.

La varianza estimada del error o cuadrado medio residual

$$MSE(p) = \frac{SSE(p)}{n - p - 1}$$

$p =$ cantidad de regresores



Se elige el modelo con menor MSE.

Equivale a elegir el modelo con mayor R^2 ajustado porque:

$$R^2_{aj} = 1 - \frac{MSE(p)}{SST/(n-1)}$$

También es usual elegir el modelo a partir de RMSE = raíz(SSE/n)

PRESS (Suma de cuadrados de Predicción)

- Supongamos que hay *p* parámetros en el modelo y que tenemos *n* observaciones disponibles para estimar los parámetros.
- En cada paso se deja de lado la *i-ésima* observación del conjunto de datos y se calcula la predicción y el residual correspondiente para la observación que no fue incluida, el cual es llamado el residual PRESS.
- Se suman todos los residuales anteriores para definer PRESS.
- Es una medida de cuán bien predice nuevos datos este modelo.
- Es usual comparar modelos con RMSECV= raíz(PRESS/n)

PRESS (Suma de cuadrados de Predicción)

$$e_{(i)} = y_i - y_{(i)} = \frac{e_i}{1-h_{ii}}$$

La medida PRESS para el modelo de regresión que contiene *p* parámetros se define por:

$$PRESS = \sum_{i=1}^{n} e_{(i)}^2 \qquad \text{o equivalentemente} \qquad PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1-h_{ii}}\right)^2$$

Se elige el modelo que tiene el valor de PRESS más bajo.

Se define también

$$R^2_{pred} = 1 - \frac{PRESS}{SST}$$

como indicador de la capacidad predictiva del modelo.

Criterio de Mallows (Cp de Mallows)

Se trata de encontrar un modelo donde *el sesgo* y *la varianza* de los valores ajustados sean moderados.

Para esto, se define el estadístico $\mathbf{C_p}$ de modo de minimizar el error cuadrático medio de un valor ajustado.

El C_p de Mallows (Draper, pag 332)

El **error cuadrático medio total normalizado** para un modelo ajustado está dado por

$$\sum_{i=1}^{n} \frac{ECMP(\hat{y}_i)}{\sigma^2} = \sum_{i=1}^{n} \frac{E[(\hat{y}_i - y_i)^2]}{\sigma^2} = \sum_{i=1}^{n} \frac{Var(\hat{y}_i) + Sesgo^2(\hat{y}_i)}{\sigma^2}$$

Se sabe que:

$$\sum_{i=1}^{n} \frac{Var(\hat{y}_i)}{\sigma^2} = traza(H) = p$$

$$E\left[\sum_{i=1}^{n} \frac{Sesgo^2(\hat{y}_i)}{\sigma^2}\right] = E[\frac{SSE}{\sigma^2}] = (n-p)$$ si el modelo es el correcto!

Criterio de Mallows (Cp de Mallows)

Se define el estadístico de *Mallows* como

$$C_p = \frac{SSE_p}{\sigma^2} + (2p - n)$$

Ojo! p = # parámetros

Se prueba que **$E(C_p)=p$** si el sesgo = 0.

Esto dice elegir un modelo con **p** parámetros tal que **C_p** sea lo más parecido posible a **p.**

SSE_p: suma de cuadrados del error del modelo que contiene ***p*** parámetros, incluyendo el intercepto,

σ^2: **varianza estimada** con el modelo completo.

Datos supervisor

Table 11.4 Values of C_p Statistic (All Possible Equations)

Variables	C_p	Variables	C_p	Variables	C_p	Variables	C_p
1	1.41	1 5	3.41	1 6	3.33	1 5 6	5.32
2	44.40	2 5	45.62	2 6	46.39	2 5 6	47.91
1 2	3.26	1 2 5	5.26	1 2 6	5.22	1 2 5 6	7.22
3	26.56	3 5	27.94	3 6	24.82	3 5 6	25.02
1 3	1.11	1 3 5	3.11	1 3 6	1.60	1 3 5 6	3.46
2 3	26.96	2 3 5	28.53	2 3 6	24.62	2 3 5 6	25.11
1 2 3	2.51	1 2 3 5	4.51	1 2 3 6	3.28	1 2 3 5 6	5.14
4	30.06	4 5	31.62	4 6	27.73	4 5	29.50
1 4	3.19	1 4 5	5.16	1 4 6	4.70	1 4 5 6	6.69
2 4	29.20	2 4 5	30.82	2 4 6	25.91	2 4 5 6	27.74
1 2 4	4.99	1 2 4 5	6.97	1 2 4 6	6.63	1 2 4 5 6	8.61
3 4	23.25	3 4 5	25.23	3 4 6	16.50	3 4 5 6	18.42
1 3 4	3.09	1 3 4 5	5.09	1 3 4 6	3.35	1 3 4 5 6	5.29
2 3 4	24.56	2 3 4 5	26.53	2 3 4 6	17.57	2 3 4 5 6	19.51
1 2 3 4	4.49	1 2 3 4 5	6.48	1 2 3 4 6	5.07	1 2 3 4 5 6	7
5	57.91	6	57.95	5 6	58.76		

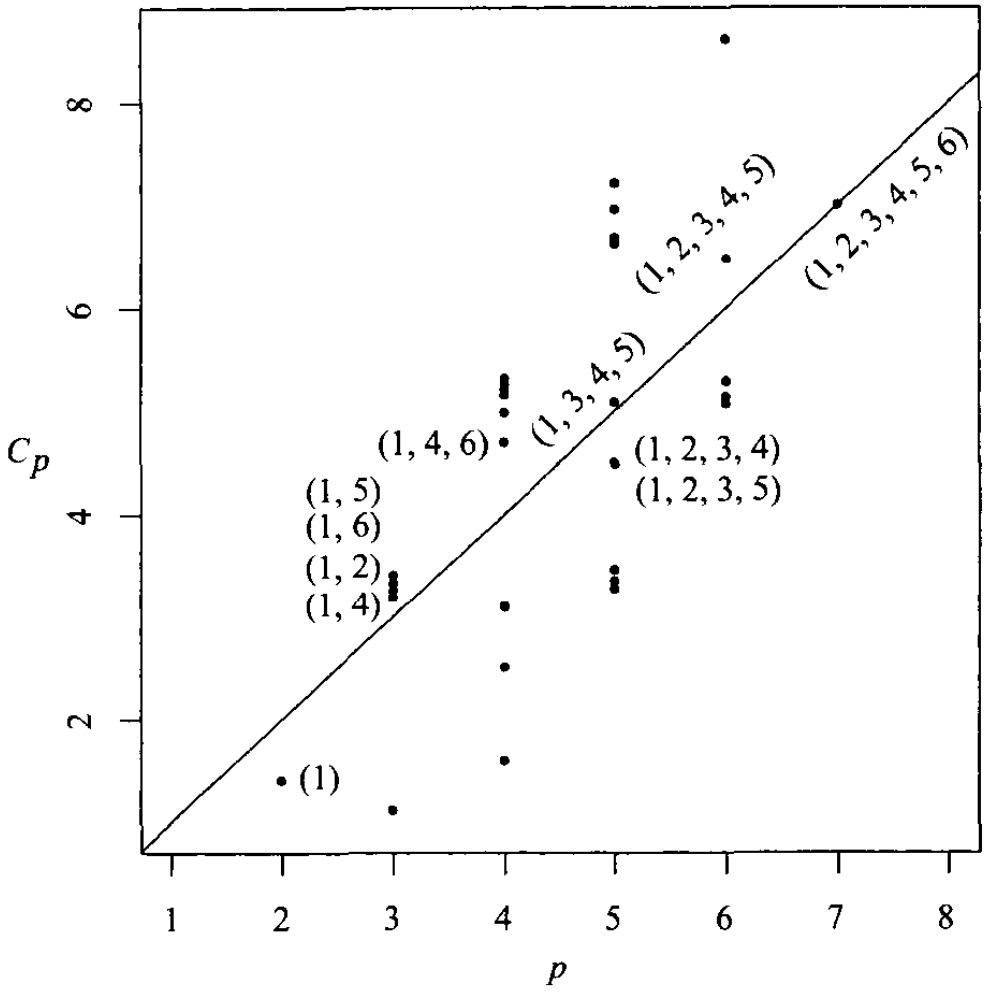


Figure 11.1 Supervisor's Performance Data: Scatter plot of C_p versus p for subsets with $C_p < 10$.

Técnicas computacionales para selección: mejor subconjunto

Requiere ajustar todas las regresiones considerando todos los subconjuntos posibles de regresores. Se evalúan todos los modelos con alguno de los criterios vistos y se selecciona el mejor modelo.

Problemas: si hay *k* regresores candidatos, hay 2^*k* subconjuntos posibles, esto es, modelos a evaluar.

Ver en (55) la comparación de todas las regresiones posibles.

Ej: datos de Hald (Montgomey)

Observation i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

TABLE 10.1 Summary of All Possible Regressions for the Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

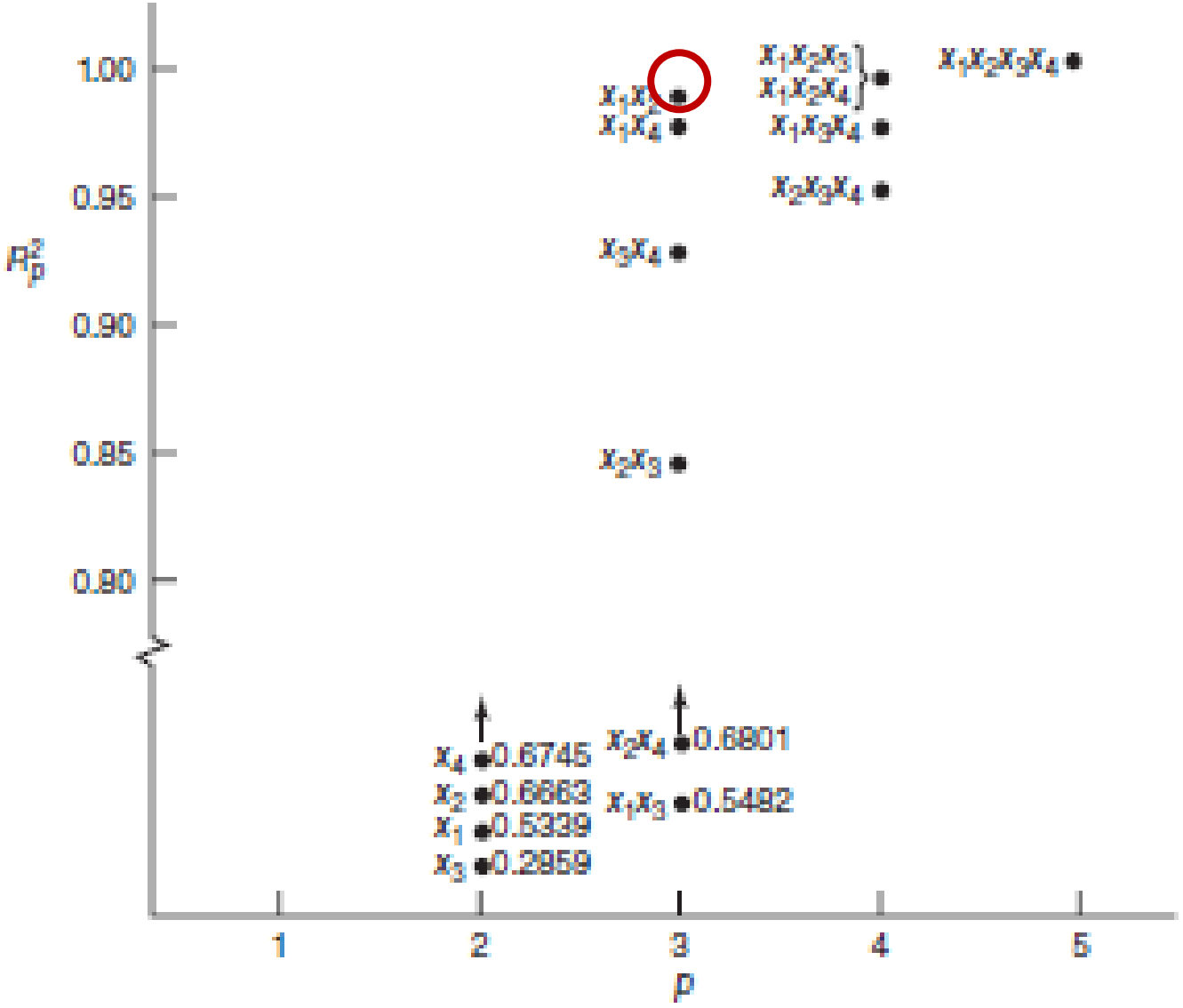


Figure 10.4 Plot of R_p^2 versus p, Example 10.1.

TABLE 10.4 Comparisons of Two Models for Hald's Cement Data

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b]		
	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	−1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	−1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	−1.5147	0.11890	2.9553	−1.8910	0.13315	4.7588
4	−1.6585	0.24225	4.7905	−1.8016	0.24431	5.6837
5	−1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	−1.3031	0.36180	4.1627	−1.4287	0.36341	5.0369
8	−2.0754	0.24119	7.4806	−3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	−2.8934	0.21420	13.5579	−2.2247	0.25923	9.0194
		PRESS x_1, x_2 = 93.8827			PRESS x_1, x_2, x_4 = 85.3516	

[a] $R^2_{\text{Prediction}} = 0.9654$, $\text{VIF}_1 = 1.05$, $\text{VIF}_2 = 1.06$.

[b] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.